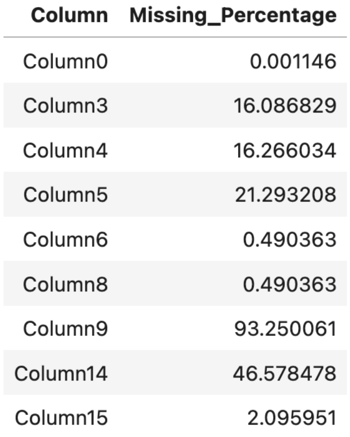
GST Challenge 2024

Introduction

The challenge involved developing a predictive model, **Fθ(X)**, capable of accurately forecasting the target variable **Y** for previously unseen data. To address this, we propose an artificial neural network (ANN)-based machine learning model. This model is designed to classify the input data into one of two categories, specifically classifying each input as either **0** or **1**, based on learned patterns and relationships within the data. The goal is to build a robust classifier that generalizes well to new instances, ensuring accurate predictions for unknown inputs.

Key methodology and steps taken

1. **Data pre-processing**: Our first step was to perform an analysis of the provided dataset, where we discovered that several columns had a significant percentage of missing values. Notably, in **Column9**, 93% of the values were missing, leading us to drop this column from the dataset. For the remaining columns with missing data, such as **Column3** (16.09%), **Column4** (16.27%), **Column5** (21.29%), and **Column14** (46.58%), we chose to impute the missing values using the respective column means to preserve the overall data structure.



Column	Missing_Percentage
Column0	0.001146
Column3	16.086829
Column4	16.266034
Column5	21.293208
Column6	0.490363
Column8	0.490363
Column9	93.250061
Column14	46.578478
Column15	2.095951

2. **Handling the class imbalance**: To address the class imbalance in the dataset, where the target distribution was heavily skewed (711,100 instances of class 0 and 74,033 instances of class 1), we applied the **SMOTE** (Synthetic Minority Over-sampling Technique) method. This helped balance the classes by generating synthetic samples for the minority class.

3. **Feature Selection**: In the feature selection process, we aimed to retain only the most relevant features for model training. To do this, we calculated the correlation between each feature and the target variable. Columns with an absolute correlation value below 0.01 were deemed to have minimal predictive power, as their relationship with the target variable was too weak. Consequently, we removed these low-correlation columns from the training set to reduce noise and improve model performance by focusing on more impactful features.

4. **Training an ANN**:

- **Model Type**: Sequential artificial neural network (ANN).
- **Input Layer**: Accepts scaled input features with shape matching the number of features in the training set.
- **Hidden Layers**:
    - **First Layer**: 128 neurons, **ReLU** activation, with **30% dropout** to prevent overfitting.
    - **Second Layer**: 64 neurons, **ReLU** activation, with **30% dropout**.
    - **Third Layer**: 32 neurons, **ReLU** activation, with **20% dropout**.
    - **Fourth Layer**: 16 neurons, **ReLU** activation, with **20% dropout**.
- **Output Layer**: 1 neuron with **sigmoid** activation for binary classification.
- **Optimizer**: **Adam** with a learning rate of **0.0001**.
- **Loss Function**: **Binary cross-entropy**, suitable for binary classification tasks.
- **Early Stopping**: Monitors validation loss with patience of 15 epochs and restores the best weights to avoid overfitting.
- **Class Weights**: Applied to handle class imbalance, with class 1 given more weight (0: 1.0, 1: 2.5).
- **Training Details**:
    - **Validation Split**: 20% of training data used for validation.
    - **Epochs**: Up to 100, with early stopping based on validation performance.
    - **Batch Size**: 64.

Model Performance Report

Test Loss: 0.09681229293346405
Test Accuracy: 0.9677355289459229

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	1.00	0.96	0.98	237034
1	0.75	1.00	0.85	24678
Accuracy			0.97	261712
Macro avg	1.00	0.96	0.98	261712
Weighted avg	0.75	1.00	0.85	261712

Confusion matrix:

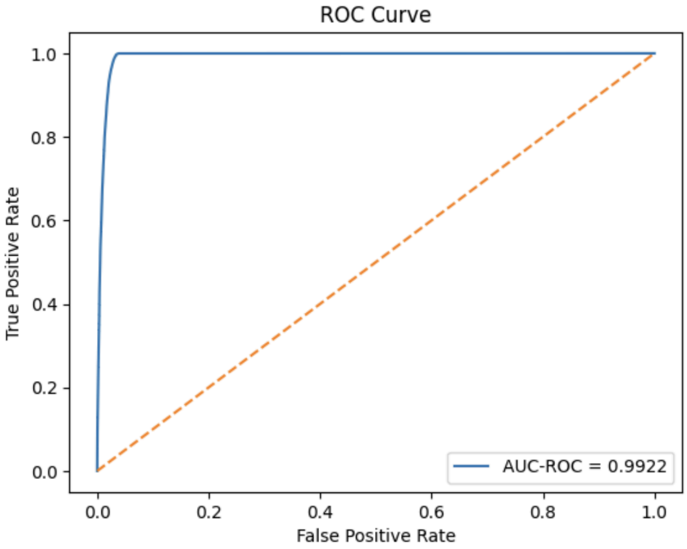
Predicted label

		0	1
Actual label	0	228627	8407
	1	37	24641

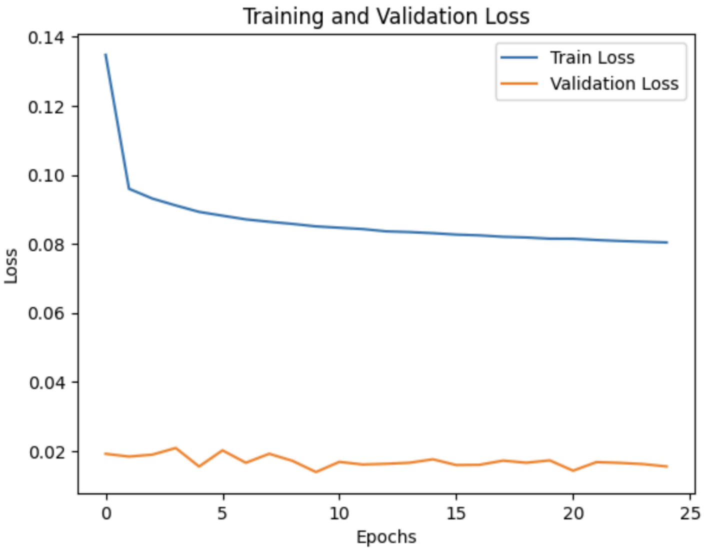
AUC-ROC curve:

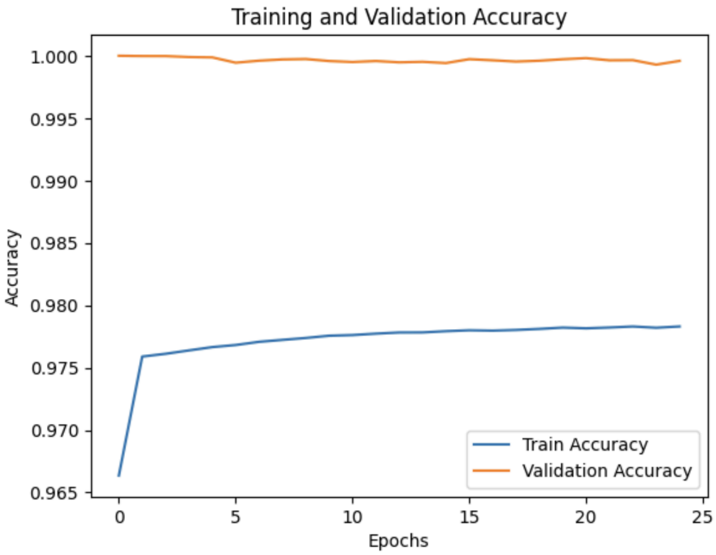
The blue line is the ROC curve for the model being evaluated. It rises very steeply and quickly reaches close to 1.0 on the y-axis, indicating excellent model performance.

The Area Under the Curve (AUC) is given as 0.9922, which is extremely close to 1. This indicates that the model has outstanding discriminative ability.



The curve's shape, being very close to the top-left corner of the plot, suggests that the model has a high true positive rate and a low false positive rate across various classification thresholds.


Training and Validation Loss
Loss
Epochs
Train Loss
Validation Loss


Training and Validation Accuracy
Accuracy
Epochs
Train Accuracy
Validation Accuracy